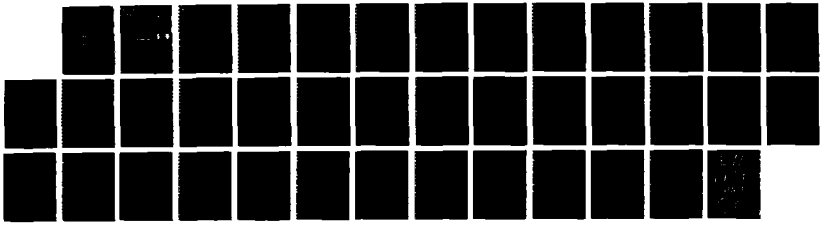
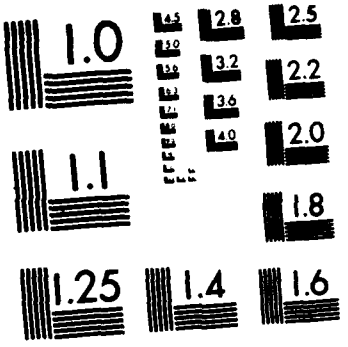
AD-A191 849 NONPARAMETRIC DISCRIMINANT ANALYSIS APPLIED TO THE EXTRACTION AND SELECTI..(U) OHIO STATE UNIV COLUMBUS ELECTROSCIENCE LAB O SNORRASON ET AL. DEC 87 1/1

UNCLASSIFIED ESL-717220-7 N00014-85-K-0321 F/G 17/9 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

AD-A191 849

(4)

THE OHIO STATE UNIVERSITY

DTIC FILE COPY

NONPARAMETRIC DISCRIMINANT ANALYSIS APPLIED TO THE EXTRACTION AND SELECTION OF RADAR TARGET SIGNATURE FEATURES

Ö. Snorrason
F.D. Garber

DTIC
ELECTE
MAR 15 1988
S D
D

The Ohio State University
ElectroScience Laboratory
Department of Electrical Engineering
Columbus, Ohio 43212

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

Final Report No. 717220-7
Contract No. N00014-85-K-0321
December 1987

Department of the Navy
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217-5000

88 3 15 024

A191 849

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO.	2.	3. Recipient's Accession No.

4. Title and Subtitle
NON PARAMETRIC DISCRIMINANT ANALYSIS APPLIED TO THE EXTRACTION AND SELECTION OF RADAR TARGET SIGNATURE FEATURES

5. Report Date
December 1987

6.

7. Author(s)
Ö. Snorrason, F.D. Garber

8. Performing Organization Rept. No.
717220-7

9. Performing Organization Name and Address
The Ohio State University
ElectroScience Laboratory
1320 Kinnear Road
Columbus, Ohio 43212

10. Project/Task/Work Unit No.

11. Contract(C) or Grant(G) No.
(C) N00014-85-K-0321
(G)

12. Sponsoring Organization Name and Address
Department of the Navy
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5000

13. Type of Report & Period Covered
Technical

14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

A nonparametric criterion of discriminant analysis is applied to a large scale data set, of radar returns from aircrafts, for frequency selection. The performance of the optimum sets characterized is evaluated using Monte-Carlo simulation. The effect of the number of frequencies and the *a priori* information of the azimuth angle on the performance of the pattern recognition system is examined. Also the effect of multiple measurements is examined.

17. Document Analysis a. Descriptors

b. Identifiers/Open-Ended Terms

c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
APPROVED FOR PUBLIC RELEASE DISTRIBUTION IS UNLIMITED	Unclassified	33
	20. Security Class (This Page) Unclassified	22. Price

(See ANSI-Z39.18) See Instructions on Reverse OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce

Contents

DTIC COPY INSPECTED 1

Accession For
NTIS CRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

List of Tables

List of Figures

I. Introduction

Reliable automatic identification of aircrafts from radar returns is of great interest in many areas of our modern society. During the last decade significant progress has been made in developing radar systems that are capable of this task, both theoretically and experimentally. For example, it has been shown, that the exact shape of an object can be reconstructed if radar return from it are available over an unlimited range of frequencies and observation angles, conditions which cannot be met in practice. It has also been shown [1], that the frequency range (Rayleigh range) corresponding to wavelengths from half the size of the object to 10 times its maximum dimension carries the essential information regarding its overall dimension and approximate shape. In practical situations it is desirable to measure the target at a limited number of frequencies to reduce the cost and complexity of the radar system. Studies have shown [2] that some frequencies are more effective for classification of aircrafts than other.

In this paper, a criterion of discriminant analysis is applied for frequency selection, to a large scale data base of radar return measurements from models of five commercial aircrafts. Our goal is to characterize the optimum sets of frequencies which minimize the classification error.

Scaled data is available for each plane at 0° elevation angle, from 0° to 180° azimuth angle in 10° steps, over a frequency range from 8 MHz to 58 MHz, in 1 MHz steps, using HHP (horizontally transmitted, horizontally received polarization) and coherent detection. Therefore, each of the 5 classes is represented by 19 prototypes, each of which is a vector of 51 complex entries (amplitude and phase). The whole measurement set has been arranged in a data base in which the measurements are scaled to square meter. All system related parameters have been removed from the data. The prototypes are considered to give exact knowledge of the classes at

corresponding angles. Once the optimum set of M frequencies has been selected using the discriminant criterion, the corresponding frequencies are extracted from the data base to produce a new smaller data base which represents the prototypes in the final feature space, the dimension of the feature space has been reduced from N (51) to $M < N$.

To estimate the performance of the pattern recognition system, the measured radar return of an unknown target is simulated. The simulation process needs to be simple enough to be implementable on a computer, but still complex enough to describe all the most important phenomena encountered in the environment of the radar system. For radar target identification systems, noise can be classified into three broad classes [3]. Noise generated by components within the radar system, noise resulting from additive (linear) sources outside the system, such as clutter and atmospheric and extraterrestial sources, and noise characterized by multiplicative disturbances, which can occur within or outside the system. Assuming ideal system, the noise is modelled as an additive white Gaussian noise on the sampled output of the output device within the radar system. This model is considered to be complex enough to represent the physical phenomena and yet parametrically simple enough to be of use in simulation and analysis. It is assumed that the noise processes at each frequency are independent identically distributed Gaussian random processes. A noisy radar return of prototype j from class l would then result in the observation, $x_m = x_{lj} + n$, where, $x_{lj} = \Re\{x_{lj}\} + j\Im\{x_{lj}\}$ and n is a vector of M complex independent identically distributed Gaussian random processes. Each component of n, n_i is generated by forming the complex sum $n_i = A_i + jB_i$, where A_i and B_i are zero mean, and $\sigma^2/2$ variance Gaussian random processes. The noise model is thus completed by the specification of the variance σ^2 of the individual complex Gaussian deviates n_i. The average noise power is specified in absolute terms in units of $\mathrm{dB}m^2$ [3].

Classification of an unknown target is done by means of the nearest neighbor (NN) decision rule. If the distance between the unknown target and one of the prototypes is shorter than that to all the other prototypes, the unknown target is given the same class identity as that closest prototype has. Thus, a classification error is declared only if wrong class is chosen.

The system performance of the selected optimum sets of frequencies is estimated at a given noise level by superimposing noise to each component of all selected frequencies. This is done for all prototypes from all the classes. Then each of these corrupted prototypes is assumed to be an unknown target and is classified according to the NN decision rule.

In the next section we will discuss the discriminant criterion used for the feature selection, and then some of the resulting optimum sets of frequencies are presented along with their classification performance curves.

II. Discriminant Analysis

In general, for multiple classes, each class of multiple prototypes, where multi-dimensional measurements are available of each prototype, it is very hard to relate a set of features to the probability of classification error. *The underlying probability distributions of the classes* must be known to find such a relationship, something which is unlikely to be available in the real world. As it is extremely difficult to find the set of features which minimizes the classification error, so called discriminant functions are widely used to obtain more reliable set of features than by picking them at random. In current litterature many different criteria have been proposed for discriminating between classes [4,5,6,7,8,9,10,11,12,13]. Criteria of this type do not utilize any information about the measurement noise. Instead they use dis-

tance measure to relate similarities within classes and resolve dissimilarities between classes, thus selecting a set of features which maximizes in some sense the average distance between the classes. Justification for the use of such criteria is based on the assumption that, the more distant the classes are at the average, the smaller is the average classification error. This assumption does though not guarantee that the selected set of features gives classification error close to the obtainable minimum using same number of features. Also, the optimum set may be dependent on the noise level. Papers have appeared, in which the optimum number of features are discussed [14,15,16,17,18,19], but there does not exist any general theory of what number of features are needed to establish some minimum average probability of error at a given noise level. Hence the number of features selected is choosen ad-hoc, mostly restricted by the computational effort needed.

In this section we describe few discriminant criteria. These criteria functions select a subspace in which the classes separate optimally in the sense of the corresponding criteria. Denote the distance, $d(x_{i,k}, x_{j,l})$ between the k-th prototype, $x_{i,k}$ from class i and the l-th prototype, $x_{j,l}$ from class j as

$$d(x_{i,k}, x_{j,l}) = (x_{i,k} - x_{j,l})^*(x_{i,k} - x_{j,l}) \quad , \tag{1}$$

where the asterisk denotes conjugate transpose.

A simple criterion function based on this distance measure for the n_i prototypes in class i, with P_i denoting the *a priori* probability of class i for $1 \leq i \leq L$ is given in [5] as

$$J_1(\mathcal{X}) = \frac{1}{2} \sum_{i=1}^{L} P_i \sum_{j=1}^{L} P_j \cdot \left(\frac{1}{n_i n_j} \right) \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} d(x_{i,k}, x_{j,l}) \quad , \tag{2}$$

which is the average probable distance between all prototypes of all L classes in the set, $\mathcal{X}$, of prototypes.

The goal of the feature selection process is to identify a subset, $\mathcal{X}$, of feature vectors in the set, $\mathbf{X}$, of possible measurement vectors so as to maximize the chosen criterion function. That is, the optimum set, $\mathcal{Z}$ of features is that which maximizes $J_1(\mathcal{Z})$, i.e.

$$\mathcal{Z} = \arg \max_{\mathcal{X} \in \mathbf{X}} J_1(\mathcal{X}) \tag{3}$$

In order to pursue the implications of the criterion function given above, notice that (2) can be written as

$$\begin{aligned} J_1(\mathcal{X}) \;\; = \;\; & \sum_{i=1}^{L} P_i \cdot (m_i - \overline{m})^*(m_i - \overline{m}) + \\ & \sum_{i=1}^{L} P_i \cdot \left(\frac{1}{n_i}\right) \sum_{k=1}^{n_i} (x_{i,k} - m_i)^*(x_{i,k} - m_i) \quad , \end{aligned} \tag{4a}$$

where m_i denotes the vector average of the n_i prototypes in class i and $\overline{m}$ denotes the vector average of all prototypes in the set $\mathcal{X}$.

The first term appearing in (4a) represents the distance of the individual class means m_i to the sample mean $\overline{m}$ and is referred to as the between class or *inter-class* distance. The inter-class distance provides a measure of the separation between the "centers" of the different prototype classes. Clearly, it is desirable to choose features so that the inter-class distance is as large as possible.

The term appearing as the second summation in (4a) represents the average of the distances of the prototypes in class i to the center or class mean m_i of class i. This distance, referred to as the average within-class or *intra-class* distance characterizes the degree of clustering of each of the L classes. In contrast to the inter-class distance, it is generally desirable to choose features so as to *minimize* the average intra-class distance.

In order to treat these terms separately, define the inter-class scatter (or covariance) matrix for the set $\mathcal{X}$ of prototypes, which is a measure of the separation between classes, as

$$S_s = \sum_{i=1}^{L} P_i \cdot (m_i - \overline{m})(m_i - \overline{m})^* \tag{5}$$

and the average intra-class scatter matrix, which is a measure of clustering within classes, as

$$S_c = \sum_{i=1}^{L} P_i \frac{1}{n_i} \sum_{k=1}^{n_i} (x_{i,k} - m_i)^*(x_{i,k} - m_i) \quad . \tag{6}$$

Then it is easy to see that the criterion function J_1 in (2) can be written as

$$J_1(\mathcal{X}) = \mathrm{tr}(S_s + S_c) \quad . \tag{7}$$

Unfortunately, this criterion function does not significantly enhance the classifiability of the prototypes in the resulting optimized feature set. In particular, notice that if either the inter-class or the intra-class distance is large, then $J_1(\mathcal{X})$ is also large; the criterion function given by (2) does not produce the desired result of minimizing the intra-class distance while maximizing the inter-class distance. Hence, the criterion function $J_1(\mathcal{X})$ gives little indication about the separability of the target classes.

A feature selection criterion function of the discriminant type that does not suffer from the shortcomings discussed above is given as in [5] by

$$J_2(\mathcal{X}) = \frac{\mathrm{tr}(S_s)}{\mathrm{tr}(S_c)} \quad , \tag{8}$$

which gives the ratio of the average inter-class distance to the average intra-class distance. This criterion function is intuitively of more utility for the radar target classification problem since this function increases as the inter-class distance increases relative to the intra-class distance, or as the intra-class distance decreases relative to the inter-class distance. Thus, maximizing this function produces a set of features such that the resulting measurement prototypes separate well into their respective classes. The principle shortcoming of the criterion function $J_2(\mathcal{X})$ is that it

is possible for correlations among the components of the prototype measurements to skew the results [5]. This situation can be rectified by preprocessing the prototype measurements.

The criterion function used for feature selection discussed in this paper is the one due to Wilks [20,21], which is based on the ratio of the "total volume" of the feature space to the clustering distance, S_c. It is given, as

$$J_4(\mathcal{X}) = \frac{\mathrm{tr}(S_t)}{\mathrm{tr}(S_c)} = \frac{1}{M}\sum_{i=1}^{M}\lambda_i \quad , \tag{9}$$

where $S_t = S_c + S_s$ is the sum of the average intra-class scatter matrix and the average inter-class scatter matrix. The covariance matrix, S_t is often referred to as the "total scatter" matrix.

In order to eliminate the possibility of skewed results due to correlations among components of the measurement vectors, $x_{i,k}$, the measurement prototypes are preprocessed so that the resulting intra-class covariances are unity, and the total scattering covariance matrix, S_t is diagonal. That is, the measurement prototypes are transformed by a mapping, $C : \mathcal{X} \longrightarrow \mathcal{Y}$ so that the components of the resulting feature vectors are uncorrelated. Thus, we form the mapping from prototypes, $x_{i,k}$ in the measurement space, $\mathcal{X}$ to prototypes $y_{i,k}$ in the feature space $\mathcal{Y}$ as

$$y_{i,k} = C\, x_{i,k} \quad , \tag{10}$$

where the matrix C is characterized by the constraints developed below.

If we denote by $S_s(y)$ the inter-class scatter matrix for the prototypes, $y_{i,k}$ in the feature space, $\mathcal{Y}$, then we see that

$$\begin{aligned} S_s(y) &= \sum_{i=1}^{L} P_i \cdot (m_y(i) - \overline{m}_y)(m_y(i) - \overline{m}_y)^* \\ &= \sum_{i=1}^{L} P_i \cdot (Cm_i - C\overline{m})(Cm_i - C\overline{m})^* \end{aligned}$$

$$= C\left\{\sum_{i=1}^{L} P_i(m_i - \overline{m})(m_i - \overline{m})^*\right\} C^*$$

or

$$S_s(y) = C\, S_s(x) C^* \tag{11}$$

Similarly, the average intra-class scatter matrix, $S_c(y)$ for the feature vectors, $y_{i,k}$ is given by

$$S_c(y) = C S_c(x) C^* \quad . \tag{12}$$

Combining these two results yields

$$S_t(y) = S_s(y) + S_c(y) = C(S_c(x) + S_s(x))C^* \quad . \tag{13}$$

Thus, we see that the requisite transformation, C, from the measurement space, $\mathcal{X}$ to the feature space, $\mathcal{Y}$ as dictated by the criterion function $J_4(\mathcal{X})$ in (9) may be found by solving the resulting set of equations for the constraints [4]:

$$C S_c(x) C^* = I \tag{14}$$

$$C S_t(x) C^* = \Lambda \quad , \tag{15}$$

where I is the $M \times M$ identity and Λ is an $M \times M$ diagonal matrix.

In order to characterize this transformation, following the procedure given in [4], we define a matrix, $S_c^{\frac{1}{2}}(x)$ such that

$$S_c^{\frac{1}{2}}(x)(S_c^{\frac{1}{2}}(x))^* = S_c(x) \quad , \tag{16}$$

and

$$S_c^{-\frac{1}{2}}(x) = (S_c^{\frac{1}{2}}(x))^{-1} \quad . \tag{17}$$

Then, we have that

$$CS_t(x)C^* = (CS_c^{\frac{1}{2}}(x))(S_c^{-\frac{1}{2}}(x)S_t(x)(S_c^{-\frac{1}{2}}(x))^*)(CS_c^{\frac{1}{2}}(x))^* \quad . \tag{18}$$

Next, we define a matrix F such that

$$F = CS_c^{\frac{1}{2}}(x) \quad , \tag{19}$$

then we have that

$$CS_c(x)C^* = CS_c^{\frac{1}{2}}(x)(CS_c^{\frac{1}{2}}(x))^* \;=\; FF^* \quad , \tag{20}$$

or from (14)

$$FF^* = I \quad . \tag{21}$$

Combining this last expression with (18) gives

$$CS_t(x)C^* = FVF^* \quad , \tag{22}$$

where V is a matrix defined by the relation

$$V = S_c^{-\frac{1}{2}}(x)S_t(x)(S_c^{-\frac{1}{2}}(x))^* \quad . \tag{23}$$

Finally, notice that the matrix, $S_c(x)$ can be written as

$$S_c(x) = EDE^* \quad , \tag{24}$$

with

$$EE^* = I \quad , \tag{25}$$

where E is a column matrix of the orthonormal eigenvectors of $S_c(x)$, and D is a diagonal matrix of the eigenvalues of $S_c(x)$. Thus, we have that

$$\begin{aligned} S_c(x) &= EDE^* = ED^{\frac{1}{2}}D^{\frac{1}{2}}E^* \\ &= ED^{\frac{1}{2}}E^*(ED^{\frac{1}{2}}E^*)^* \end{aligned}$$

and

$$S_c^{-1}(x) = S_c^{-\frac{1}{2}}(x)(S_c^{-\frac{1}{2}}(x))^* \quad . \tag{26}$$

This gives

$$S_c^{\frac{1}{2}}(x) = ED^{\frac{1}{2}}E^* \tag{27}$$

and

$$S_c^{-\frac{1}{2}}(x) = ED^{-\frac{1}{2}}E^* \quad . \tag{28}$$

As a result, we see that the transformation matrix C is found as follows:

1. Find the decomposition

$$S_c(x) = EDE^* \quad \text{where} \quad EE^* = I \quad . \tag{29}$$

2. Find $S_c^{-\frac{1}{2}}(x)$ using the relation

$$S_c^{-\frac{1}{2}}(x) = ED^{-\frac{1}{2}}E^* \quad . \tag{30}$$

3. Calculate the matrix V as

$$V = S_c^{-\frac{1}{2}}(x)S_t(x)(S_c^{-\frac{1}{2}}(x))^* \quad . \tag{31}$$

4. Find the matrix F that satisfies

$$FVF^* = \Lambda \quad \text{or} \quad V = F^*\Lambda F \quad , \tag{32}$$

where $FF^* = I$.

5. Calculate the transformation matrix C as

$$C = FS_c^{-\frac{1}{2}}(x) \quad . \tag{33}$$

Thus, we see from (33) that the transformation C from the measurement space $\mathcal{X}$ to the feature space $\mathcal{Y}$ consists of a weighting transformation, $S_c^{-\frac{1}{2}}(x)$, and an orthonormal transformation, F.

The weighting transformation, $S_c^{-\frac{1}{2}}(x)$ is obtained as the "inverse square-root" of the average intra-class scatter matrix. This portion of the transformation matrix C has the effect of minimizing the average intra-class distance.

The second component matrix of the transformation given by (33) is, by definition, a unitary transformation and, as such, merely performs a rotation operation. In particular, the matrix, F is chosen so that average intra-class scatter matrix for the feature vectors is the identity matrix and the sum of the average intra-class and inter-class scatter matrices is a diagonal matrix after the transformation.

In summary, we have seen that the transformation matrix, C maps the measurement space, $\mathcal{X}$ into a feature space, $\mathcal{Y}$ where the prototype vectors, $y_{i,k}$ have intra-class covariances that are evenly distributed and inter-class covariances that are zero between classes. Since the transformed prototypes, $y_{i,k}$ from different classes are uncorrelated, the separability of the resulting set of prototypes can be considered on a component-wise basis. From (15) we see that the contributions of the feature vector components to the optimization are additive. Hence, if it is desired to find the optimal set of M features for the set of targets of interest, it is necessary only to find those M component features that individually provide maximum separation between the target classes in the sense of the criterion, J_4 given in (9).

Notice from (9) that each λ_k, $k = 1, \cdots, M$ is an eigenvalue of the matrix $S_c^{-\frac{1}{2}}(x)S_t(x)(S_c^{-\frac{1}{2}}(x))^*$. Also notice that

$$S_c^{-\frac{1}{2}}(x)S_t(x)(S_c^{-\frac{1}{2}}(x))^* = S_c^{-\frac{1}{2}}(x)S_s(x)(S_c^{-\frac{1}{2}}(x))^* + I \tag{34}$$

Now by forming the transformation matrix, Q, such that $QQ^* = I$ and

$$Q\left[S_c^{-\frac{1}{2}}(x)S_s(x)\left[S_c^{-\frac{1}{2}}(x)\right]^* + I\right]Q^* \;=\; \tilde{\Lambda} + I \quad , \tag{35}$$

where $\tilde{\Lambda}$ is an $M \times M$ diagonal matrix of the eigenvalues, $\tilde{\lambda_k}$, of $S_c^{-\frac{1}{2}}(x)S_s(x)(S_c^{-\frac{1}{2}}(x))^*$, then we have that [5]

$$\lambda_k = \tilde{\lambda_k} + 1 \quad . \tag{36}$$

Since, by definition, $\tilde{\lambda_k} \geq 0$, then the minimum value of λ_k is 1.

Intuitively, we see that each eigenvalue, λ_k gives the ratio of the sum of the average inter-class and intra-class distances to the average intra-class distance, in the direction of the corresponding eigenvector. Thus, the constraint $CS_cC^* = I$ imposed on the mapping C has the effect of "normalizing" the distance measure in the feature space, $\mathcal{Y}$ so that each of the eigenvalues give an absolute indication of how well the classes separate in the corresponding direction. If, for example, it happens that λ_k is large for a certain value of k then we would expect the classes to separate well when the k-th component feature is employed for classification. On the other hand, if a feature component has an eigenvalue of 1, then the average inter-class distance is no larger than the average intra-class distance in that direction. This, in turn, implies that the target classes are not separable in the direction of the corresponding eigenvector.

The criterion function, J_4 of (9) can be employed either as the basis for feature selection, or as the discriminant function for the feature extraction process. If the criterion function is used for feature selection, then F becomes an $M \times M$ matrix, with row vectors that are the eigenvectors of the M-dimensional subspace producing the largest sum of M eigenvalues. Thus, in this case, the transformation matrix, C is $M \times M$.

If, on the other hand, the function, J_4 is used as a criterion for feature extraction,

then F becomes an $M \times N$ matrix, with row vectors that are the eigenvectors of the corresponding M largest eigenvalues of the $N \times N$ matrix $S_c^{-\frac{1}{2}}(x)S_s(x)(S_c^{-\frac{1}{2}}(x))^*$. In this case, C becomes a matrix with dimension $M \times N$.

Finally, we point out that while several other discriminant-based algorithms for feature selection and extraction have been proposed [6,22,23,12], these criteria place little emphasis on distributing the total interclass distance equally among the classes in the resulting coordinate system. In contrast, it is unlikely that a feature selection or extraction process based on J_4 would result in good separability between two classes in the target set, at the expense of separation between other classes in the set [5]. In addition, the criterion J_4 is moderately easy to implement and possesses many of the properties desirable in a discriminant function. The performance of J_4 should give a good indication of how powerful tool discriminant analysis is for feature selection for composite classes.

Most often, discriminant analysis is applied to simple classes, where noisy training samples are available for each class. In this case, the scatter of the prototypes is due to the measurement noise instead of the spread of many exact prototypes of composite classes. Generally speaking; for simple classes, where only noisy training samples are available, the discriminant algorithm selects a subspace where the noise level is low compared to the average inter-class distance; for composite classes, where exact information of the subclasses is available, the discriminant algorithm selects a subspace where the average inter-class distance is large compared to the average intra-class distance. For a simple class, the mean of all the noisy prototypes (training samples) gives moderately good representative mean vector for the class. On the other hand, if the class is composite, the mean vector of the class prototypes may not be a good representative for the class. This may be the most severe shortcoming of applying discriminant analysis methods to composite classes.

III. Results

The discriminant criterion described above was applied to the data set assuming either no prior information about the azimuth angle of the target, or assuming $\pm 20°$ prior directional information. The optimum set of M frequencies is found by selecting every possible subset of M frequencies out of N (51), calculate the criterion value and select the subset which produces highest criterion value. As the desired number of frequency measurements increases (up to 26), the number of possible subsets increases drastically. Also, as the number of features is increased the computation for each subset takes more time and becomes very exhaustive for ordinary computer system in terms of cpu time.

A. No Prior Information of the Azimuth Angle

For coherent detection the 10 best sets of 1, 2, 3, 4 and 5 frequencies were obtained. The result of this search for the optimum set of 4 frequencies is shown in Table 1. There is less than 1% difference between the criteria value of the optimum set and the value of the 10.th best set. This indicates low sensitivity of exactly which set of frequencies is selected as the optimum one. If simulation is done using the optimum set and the 10.th best set, the classification result, as expected, is very similar. If the number of desired frequencies exceeds 5 the search becomes extremely exhaustive and time consuming. By comparing the selection result for the 10 best sets of frequencies for the desired number of frequencies being 3, 4 and 5, there is high tendency to select some of the same frequencies all the time. Hence to make the search less exhaustive and still be able to obtain higher number of desired frequencies, 4 of the frequencies were choosen beforehand and the feature selection algorithm then used to search those 4 more frequencies which would maximize the criteria function, resulting in the "optimum" 8 frequencies. The optimum sets of

1, 2, 3, 4, 5 and "8" frequencies are shown in Table 2. There is, of course, no guarantee that this set of 8 frequencies is the optimum set in the sense of the criteria function, but very likely they are close to the optimum 8, especially as there is low sensitivity of which set of frequencies it the optimum one. Figure 1 shows the classification result using the optimum set of 4 frequencies and a set of 4 equally spaced frequencies. Clearly, the optimum set yields better performance than the set of equally spaced frequencies, though the improvement is not large. Similar results were also obtained for corresponding sets of 3, 5, and 8 frequencies.

In Figure 2 the classification curves for the optimum set of 3, 4, 5 and "8" frequencies are compared to the curve obtained when using all the 51 frequencies. It is obvious that the classification result is much better if all the 51 frequencies are used. This is not very surprising as all the eigenvalues found in the feature selection algorithm were approximately of the order 1.0 - 1.8, which means that the separation in any direction in the feature space is of the same magnitude and the classes are close to each other and heavily overlapping. That might have been expected as the azimuth angle varies from 0° to 180°, causing a huge change in the effective area of the planes. For 20% probability of misclassification, the 5 optimum frequencies can operate in around 22 dBm^2 noise level, while all the 51 frequencies can operate in around 32 dBm^2 noise level. If the target were now measured at the 5 best frequencies 10 times, altogether requiring 50 measurements, and the average of these 10 sets of measurements used for the classification algorithm (SNR in the measured signal has been increased by 10 dB), the performance of this classification scheme is just sligthly better than the performance of the classifier using all the 51 frequencies measured once. This implies it is sufficient to measure the target only at the optimum set of 5 frequencies 10 times and still have about the same information about the target as if it were measured at all the 51 measurements once. The classification result is shown in Figure 3. In Figure 4 it is shown how

much the improvement is if the measurement of the set of 5 optimum frequencies is increased by one at the time. The largest improvement, by adding a measurement of the best set, is when the best set is measured twice instead of once. As the number of measurements is increased the less impact has each additional set of measurement on the improvement of the classifier, i.e. the improvement levels off. The same tendency can also be observed on Figure 2 for the number of frequencies. There is about the same improvement going from 4 frequencies to 5 frequencies as going from 5 frequencies to 8. In other words the performance advantage becomes less and less significant as the number of frequencies is increased or the number of multiple measurements is increased.

B. Partial ±20° Information of the Azimuth Angle

Modern radar system can be used to give information about the direction of the target. Hence it is realistic to consider pattern recognition system were some prior directional information of the target is available.

Assuming prior information of the azimuth angle of the target to be within ±20° uncertainity, the best sets of 4 frequencies were found for each azimuth angle. This is done by finding which frequencies discriminate the classes best at each azimuth angle, ±20°. This will result in 19 sets of optimum frequencies, one set for each azimuth angle. Now if an unknown target were being measured, its approximate direction (azimuth angle) would need to be estimated and then the target would be measured at the optimum set of frequencies selected for that azimuth angle and classified by comparing it to the corresponding noise free prototypes. The optimum sets of 4 frequencies found at every azimuth angle are shown in Table 4. It is interesting to see that comparatively lower frequencies are selected to discriminate the airplanes at the broadside than at azimuth angles closer to 0° or 180°. The physical interpretation of this could be that at the broadside, lower frequencies

carrying size information give best discrimination, while higher frequencies carrying information about more details give best discrimination at 0° and 180°. It was observed that if the test set were loaded with prototypes at azimuth angle $i° + 20°$ and the catalogue set with prototypes at azimuth angle $i°$, $i° \pm 10°$ and $i° \pm 20°$ at the frequencies selected for azimuth angle $i°$, that the classification performance did not degrade much from what is was if the test set contained prototypes at $i°$ azimuth angle. This means that if a target is estimated to be at $i°$ but is in reality within $i \pm 20°$ the classification algorithm is not sensitive because of this inaccuracy in the azimuth angle. However, this does not state anything about the performance of the system if the target is not at an aspect angle which is not a multiple of 10°.

To examine what is gained by having ±20° prior information of the azimuth angle, the test set was loaded with prototypes at angle i, at the opimum 4 frequencies found when no prior information existed, and the test set classified to the whole catalogue containing every azimuth angle. Then the test set were loaded with the optimum 4 best frequencies at angle $i°$ given ±20° knowlegde of the angle and the catalouge set loaded with the same frequencies at angles $i°$, $i° \pm 10°$ and $i° \pm 20°$. On Figures 5 - 7 comparison of the classification curves for the classifier with no azimuth information and the one with ±20° prior information is shown for 0°, 90° and 180°. Except for 90°, the performance of the classifier with the prior information is superior to the performance of the one with no information. The reason for this bad result at the broadside might be that there is largest difference between the measurement value for the huge airplanes and the small ones, hence the selection algorithm favours just the discrimination between the large and the small groups, resulting in bad discrimination within each size group. At 90° the optimum set was selected as the four lowest frequencies. But overall, there is significant improvement in the classification performance given the prior information of the azimuth angle. This can be due to two factors:

- the optimum frequencies for each azimuth angle do give better discrimination between the classes at corresponding azimuth angle than the optimum set of 4 frequencies do, when no prior information existed.

- the improvement is due to the decreased number of prototypes in the catalogue set for each class (5 prototypes instead of 19).

One would expect the improvement to be be due to both of these factors. To find how much is gained by each factor, prototypes at azimuth angle i were loaded into the test set, and prototypes at azimuth angle i, $i \pm 10°$ and $i \pm 20°$ were loaded into the catalogue set at the optimum 4 frequencies found assuming no prior information about the azimuth angle. On Figure 8 the classification curves for targets at 30° are compared to each other, using the optimum 4 frequencies when no prior information exists, the optimum 4 frequencies with ±20° prior information and the optimum 4 frequencies when no prior information exists but only using prototypes within ±20° from the real azimuth angle in the catalogue set. Results were also obtained for every 30° azimuth angle. It was observed, that at the average the improved performance when prior azimuth angle information exist is equally due to the prior information of the azimuth angle and the decrease of the number of prototypes in each class from 19 to 5.

IV. Discussion

The classification improvement using the optimum set of frequencies compared to the set of equally spaced ones is rather low. There are probably many reasons for this. The criterion function used to select the frequencies does not bear any direct relation to the probability of misclassification criterion. It only finds the frequencies

at which the classes separate best in a noiseless environment, hopefully implying that the farther away the classes are at the average, the less is the probability of classification error. When no prior information of the azimuth angle existed the criteria value is rather low. This indicates that there is not large separation of the classes and they probably overlap each other heavily. This should be expected, as the azimuth angle is changed from 0° to 180°, the effective area of the targets changes very much. Hence, a small airplane on the broadside may look similar to a large airplane under some other angle. Also, there is small variation in the criteria value for the 10 best sets of frequencies. This suggests that there does not exist a set of frequencies which can discriminate the planes much better than any other set. The 5 classes form 10 different pairs of classes, all of which needs to be separated from each other. The feature selection algorithm were used for each pair of classes to obtain the optimum 4 frequencies which separate corresponding pair of classes best. The result can be found in Table 3 which includes the optimum 4 frequencies for all the classes. Table 3 shows that none of the optimum 4 frequencies for some of the pairs is in common with any of the optimum 4 frequencies for all the classes. Some other pairs have only one frequency in common with the overall optimum 4 frequencies. This would suggest that separation between corresponding pairs of classes is poor and worse than between pairs of classes which have more number of optimum frequencies in common. Probably the main weakness of this criteria function is most obviously exposed when it selected the optimum 4 frequencies for $\pm 20°$ prior information at 90° azimuth angle. At this angle the area difference between the large and small airplanes is greatest, giving huge separation between the group of large planes and the group of small planes, but not taking into account the separation between airplanes within each group, i.e. the criteria function favours large separation between some classes at the cost of small separation between other classes. In Table 3 it can be seen that for 90° azimuth angle the 4 lowest frequencies

were selected, all of which carry mostly size information.

In summary, the optimum sets of frequencies characterized by the discriminant criterion yield in general to lower classification error than the corresponding sets of equally spaced frequencies.

Table 1: Discriminant analysis, optimum sets of 4 frequencies, coherent measurements, HHP.

selected frequencies [MHz] 8.....20.....30.....40.....50.....58	criteria value	Pave [dBm^2]
..11.....1.....1	4.75	26.0
..11.....1.....1	4.74	26.1
..11.....1.....1	4.73	26.1
..11.....1.....1	4.70	26.1
11.....1.....1	4.70	26.6
..1.1.....1.....1	4.69	25.9
..1.1.....1.....1	4.69	25.9
1.11.....1	4.68	26.7
11.....1.....1	4.68	26.6
..1.....1.....1.....1	4.67	25.9

Table 2: Discriminant analysis, optimum sets of 1,2,3,4,5 and 8 frequencies, coherent measurements, HHP.

selected frequencies [MHz] 8.....20.....30.....40.....50.....58	criteria value	Pave [dbm^2]
.....1	1.18	22.3
.....1.....1	2.36	24.5
..1.....1.....1	3.55	25.8
..11.....1.....1	4.75	26.0
1.11.....1.....1	5.92	26.5
1.11.....1.....1..1.....11	9.30	26.3

Table 3: Discriminant analysis, optimum set of 4 frequencies every pair of airplanes, coherent measurements, HHP.

selected frequencies [MHz] 8.....20.....30.....40.....50.....58	criteria value	Pave [dBm^2]	pair of planes
.....1.11.....1.....	4.60	24.6	707-727
..1.....1.....1.....1	5.38	26.4	707-747
1.1.....11.....	4.66	26.3	707-DC10
..1.....11.....1.....	4.59	24.8	707-CON
..11.....1.....1	5.31	26.2	727-747
.1111.....	4.66	28.0	727-DC10
1111.....	4.86	26.6	727-CON
....1.....1.....1.....1	4.92	26.5	747-DC10
.....1.....1.....1.1	4.95	28.3	747-CON
1.....1.....1.....1....	4.36	25.1	CON-DC10
..11.....1.....1	4.75	26.0	Optimum 4

Table 4: Discriminant analysis, azimuth dependent optimum sets of 4 frequencies, $\pm 20°$ partial azimuth information, HHP.

selected frequencies [MHz] 8.....20.....30.....40.....50.....58	criteria value	Pave [dBm^2]	azimuth angle
..11...1.1.....	30.99	24.4	0
1.....11.1	12.10	22.9	10
11.....11.....	8.05	26.4	20
.....1..1.....1.....1	8.56	21.9	30
.....11....1.....1	9.54	19.8	40
.....1....1.....11.....	9.01	25.7	50
.....1.....1....1....1...	8.09	25.8	60
.....111.....1.....	7.24	29.2	70
.11.....1.....1.....	8.11	29.8	80
1111.....	10.38	29.2	90
11.....1....1.....	10.95	29.5	100
111.....1.....	12.49	28.5	110
11.1.....1.....	10.85	26.9	120
.....1.....1.1...1.....	9.15	22.5	130
1..1.....1.....1.....	9.07	25.3	140
1.1.....1.....1.....	10.26	26.1	150
1.1..1.....1.....	13.21	26.4	160
.1.....1.....1.....1.....	22.55	24.3	170
1.....1.....1.....1....	109.3	25.0	180

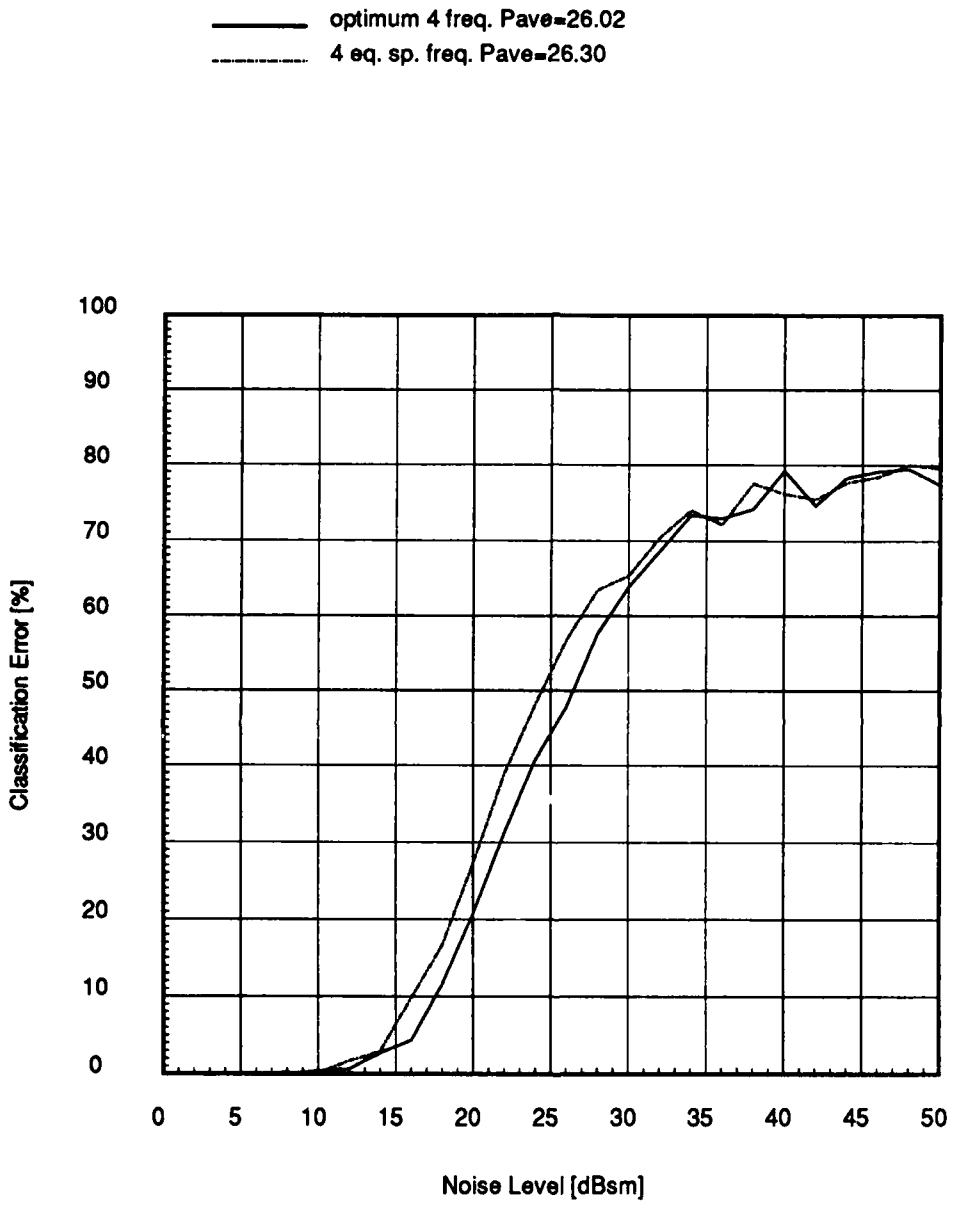


Figure 1: Performance of the optimum set of 4 frequencies and the set of 4 equally spaced frequencies, no prior information.

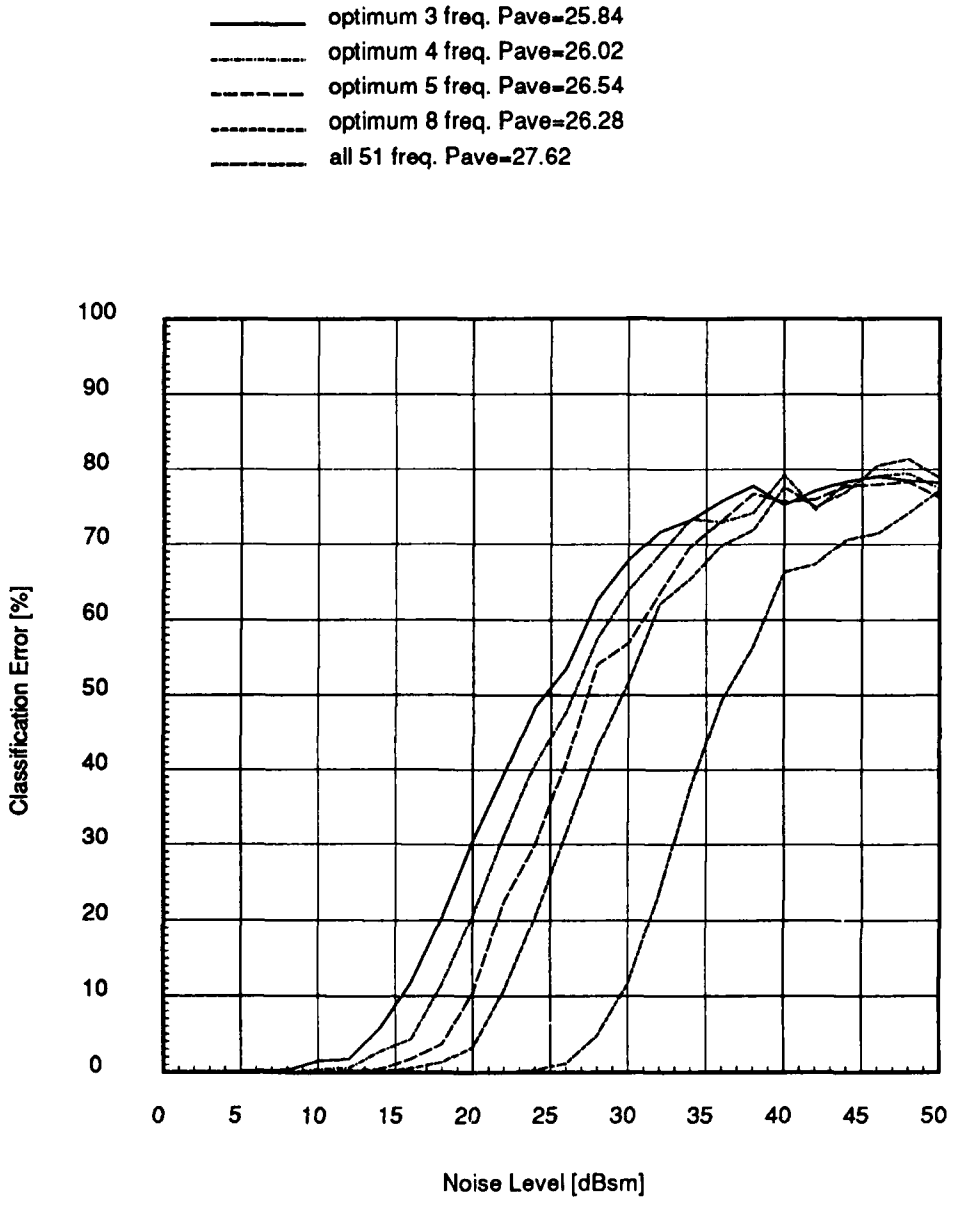


Figure 2: Performance of the optimum sets of 3, 4, 5 and "8" frequencies and the whole measurement set of 51 frequencies no prior information.

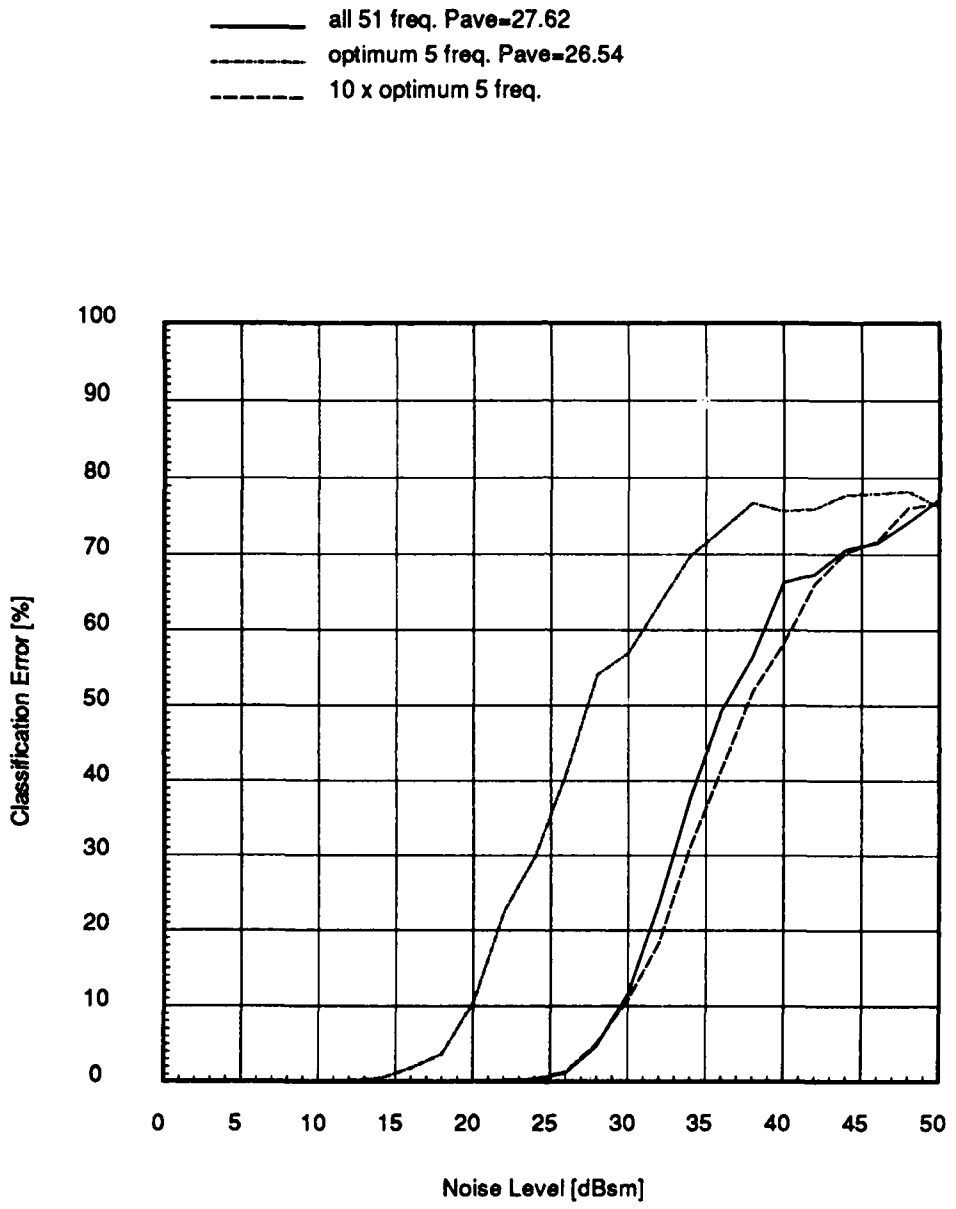


Figure 3: Performance of the optimum set of 5 frequencies, the optimum set of 5 frequencies measured 10 times and the whole measurement set of 51 frequencies, no prior information.

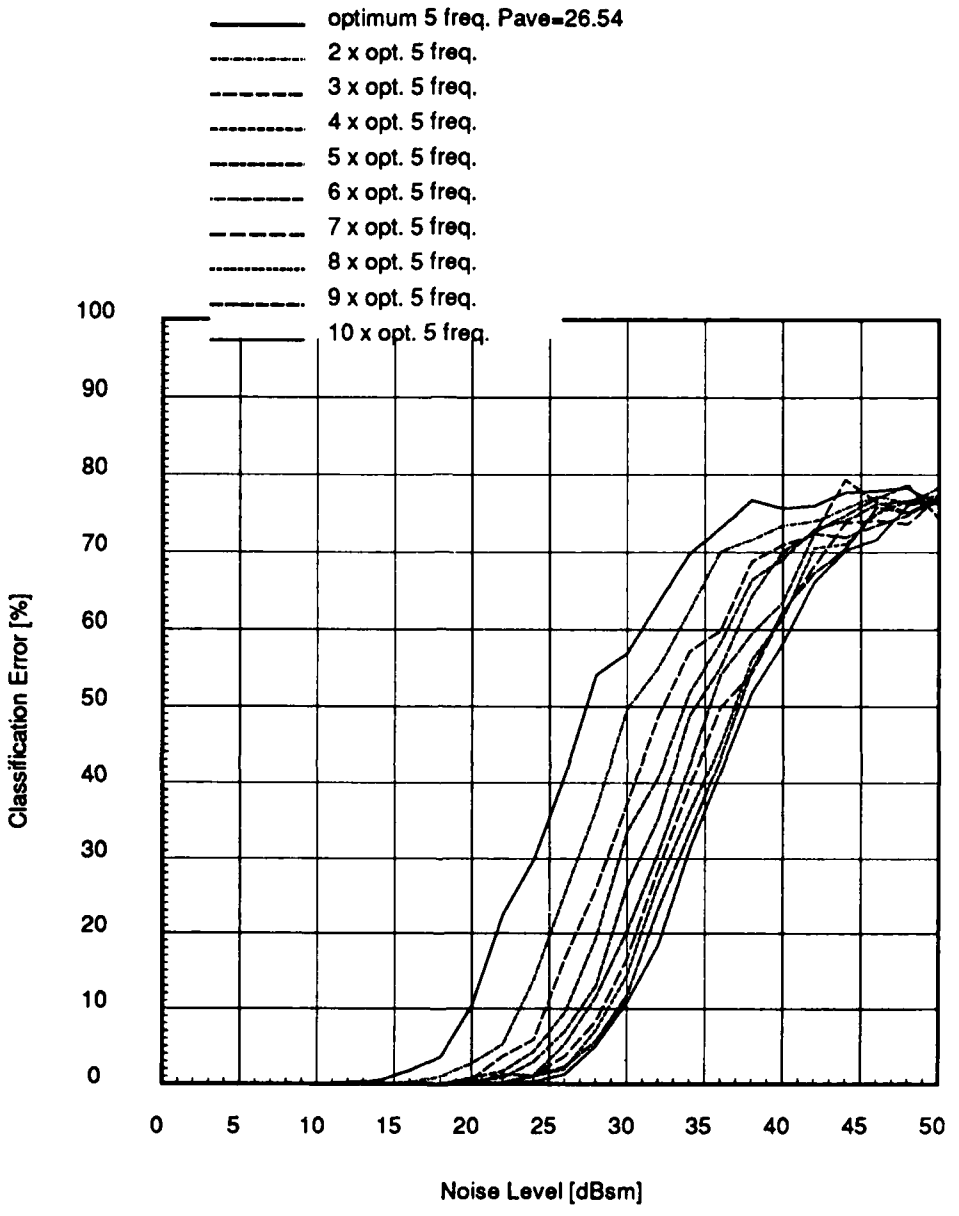


Figure 4: Performance of the optimum set of 5 frequencies measured 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 times, no prior information.

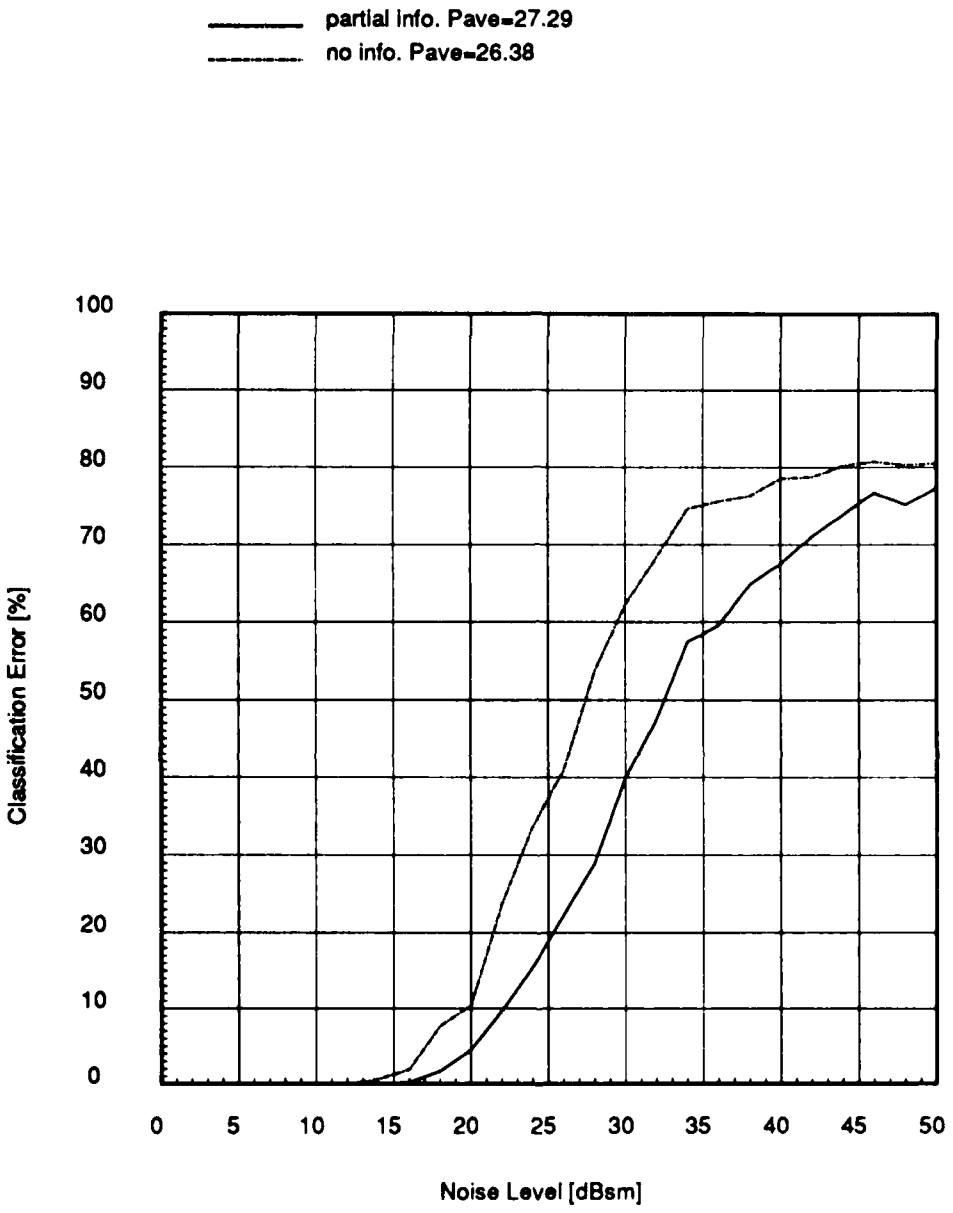


Figure 5: Performance of the optimum set of 4 frequencies assuming $\pm 20°$ prior information and the optimum set of 4 frequencies assuming no prior information, target at 0°.

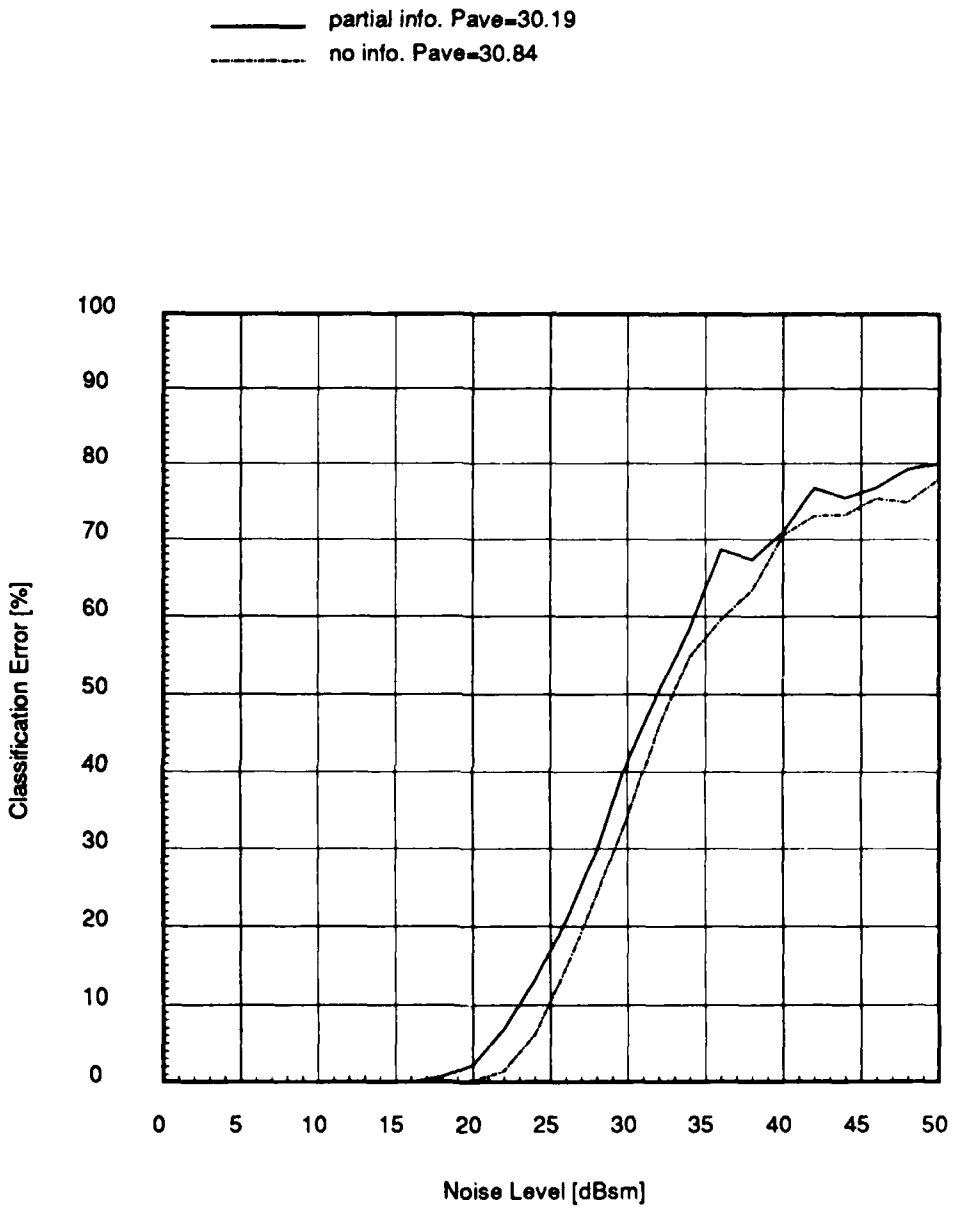


Figure 6: Performance of the optimum set of 4 frequencies assuming $\pm 20°$ prior information and the optimum set of 4 frequencies assuming no prior information, target at 90°.

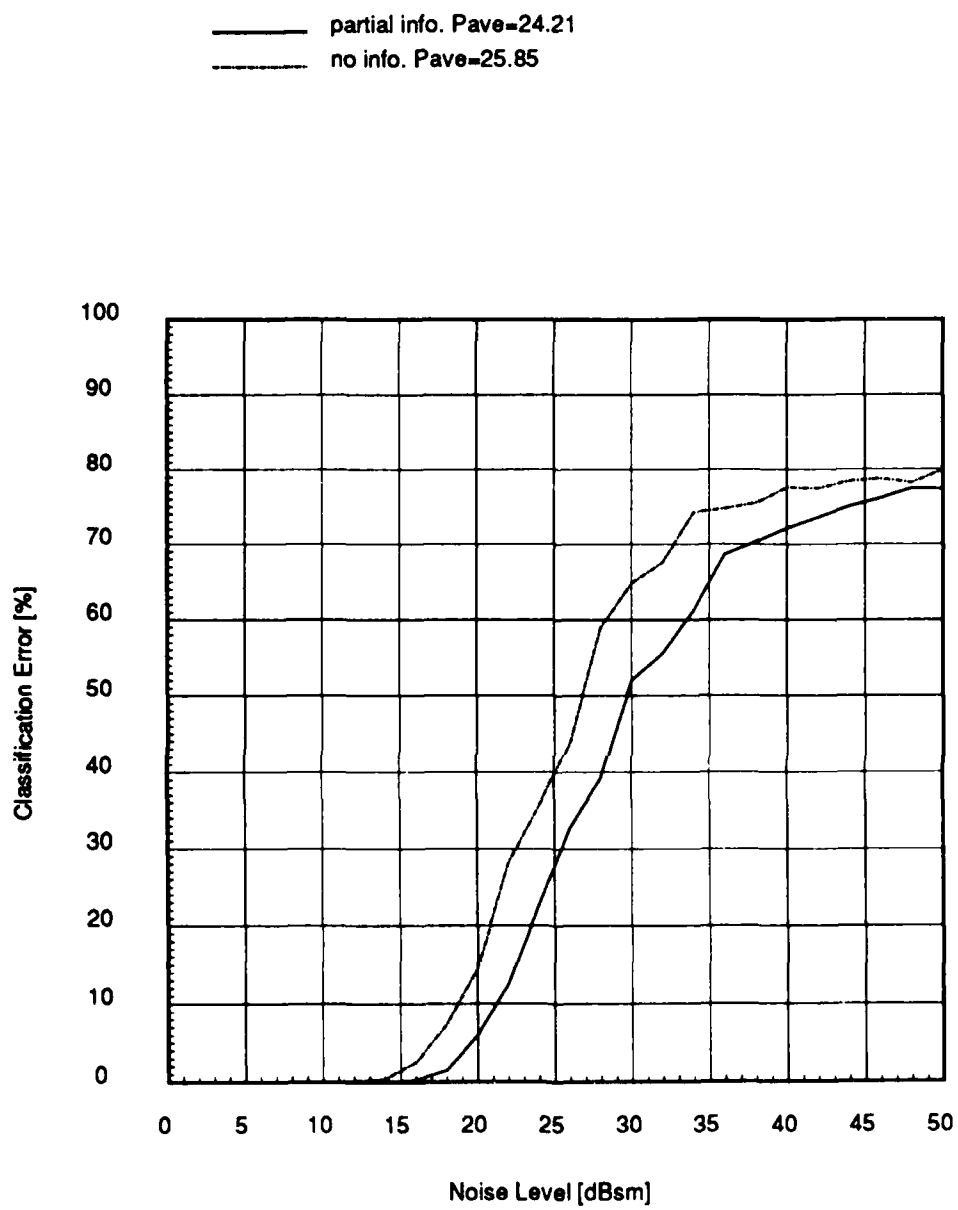


Figure 7: Performance of the optimum set of 4 frequencies assuming $\pm 20°$ prior information and the optimum set of 4 frequencies assuming no prior information, target at 180°.

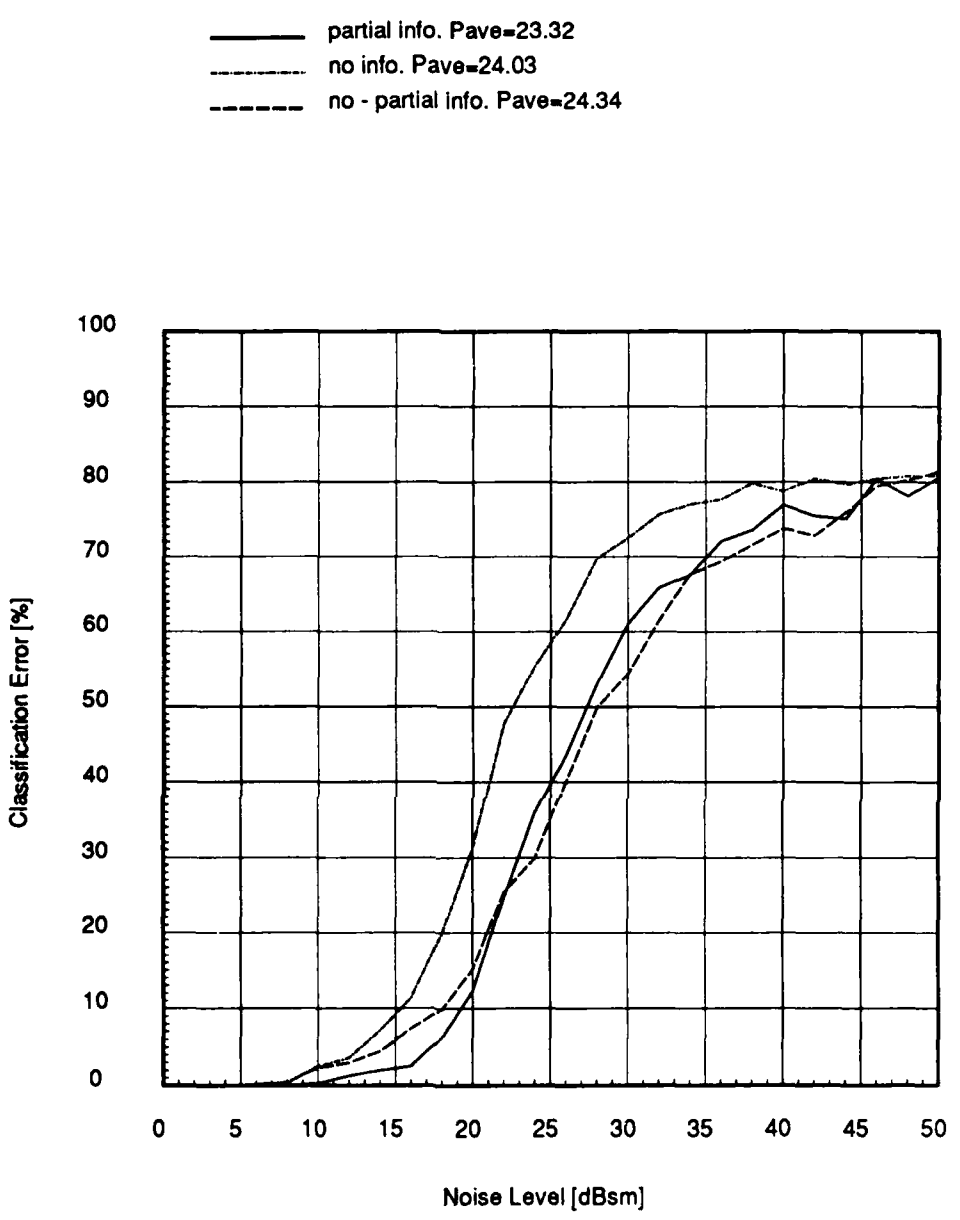


Figure 8: Performance of the optimum set of 4 frequencies assuming $\pm 20°$ prior information, the optimum set of 4 frequencies assuming no prior information but classified assuming $\pm 20°$ information, and the optimum set of 4 frequencies assuming no prior information, target at 30°.

References

[1] E. M. Kennaugh and D. L. Moffatt, "Transient and inpulse response approximation," *Proceedings of the IEEE*, vol. 53, no. 8, pp. 893–901, August 1965.

[2] H. Lin and A. A. Ksienski, "Optimal frequencies for aircraft identification," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES–17, no. 5, pp. 656–665, September 1981.

[3] F. D. Garber, "Applications of compact-range data to radar system simulation and evaluation," in *1985 Workshop on Measurement, Processing, and Analysis of Radar Target Signatures*, The Ohio State University, Columbus, Ohio, September 10–13, 1985.

[4] S. Bow, *Pattern Recognition.* New York: Dekker, 1984.

[5] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* New York: Prentice Hall, 1982.

[6] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1972.

[7] K. Fukunaga and J. M. Mantock, "Nonparametric discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI–5, no. 6, pp. 671–678, November 1983.

[8] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Princilpes.* Reading, Massachusetts: Addison–Wesley, 1974.

[9] J. Bryant and L. F. Guseman, Jr., "Distance preserving linear feature selection," *Pattern Recognition*, vol. 11, pp. 347–352, 1979.

[10] H. P. Decell and L. F. Guseman, Jr., "Linear feature selection with applications," *Pattern Recognition*, vol. 11, pp. 55–63, 1979.

[11] E. M. Rounds, "A combined nonparametric approach to feature selection and binary decision tree design," *Pattern Recognition*, vol. 12, pp. 313–317, 1980.

[12] J. Van Ness, "On the prominence of non-parametric Bayes rule discriminant algorithms in high dimensions," *Pattern Recognition*, vol. 12, pp. 355–368, 1980.

[13] K. A. Brakke, J. M. Mantock, and K. Fukunaga, "Systematic feature extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI–4, no. 3, pp. 291–297, May 1982.

[14] A. K. Jain and W. G. Waller, "On the optimal number of features in the classification of multivariate Gaussian data," *Pattern Recognition*, vol. 10, pp. 365–374, 1978.

[15] S. J. Raudys, "Determination of optimal dimensionality in statistical pattern classification," *Pattern Recognition*, vol. 11, pp. 263–270, 1979.

[16] P. L. Odell, "A model for dimension reduction in pattern recognition using continuous data," *Pattern Recognition*, vol. 11, pp. 51–54, 1979.

[17] T. S. El-Sheikh and A. G. Wacker, "Effect of dimensionality and estimation on the performance of Gaussian classifiers," *Pattern Recognition*, vol. 12, pp. 115–126, 1980.

[18] P. Pudil and S. Blaha, "Evaluation of the effectiveness of features selected by discriminant analysis methods," *Pattern Recognition*, vol. 13, pp. 81–85, 1981.

[19] S. D. Morgera and L. Datta, "Toward a fundamental theory of optimal feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI–6, no. 5, pp. 601–616, May 1984.

[20] S. S. Wilks, *Mathematical Statistics.* New York: Wiley, 1962.

[21] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[22] W. Malina, "On an extended Fisher criterion for feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, no. 5, pp. 611–614, September 1983.

[23] C. Peters, "Feature selection for the best mean square approximation of class densities," *Pattern Recognition*, vol. 11, pp. 361–364, 1979.

END

DATE

FILMED

5-88

DTIC